DOCUMENT RESUME

ED 279 708 TM 870 120

AUTHOR Wainer, Howard; Kiely, Gerard L.
TITLE CATs, Testlets, and Test Construction: A Rationale for Putting Test Developers Back into CAT.
INSTITUTION Educational Testing Service, Princeton, NJ. Program Statistics Research Project.
REPORT NO ETS-RR-86-33; ETS-TR-86-71
PUB DATE Sep 86
NOTE 46p.; Part of the Program Statistics Research Technical Report Series.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; Algorithms; *Computer Assisted Testing; Construct Validity; Difficulty Level; *Item Banks; Latent Trait Theory; Mathematical Models; Measurement Techniques; Psychometrics; *Test Construction; *Test Format; Testing Problems; Test Items; Test Theory; *Test Validity
IDENTIFIERS *Test Developers; Unidimensionality (Tests)

ABSTRACT

Recent experience with the Computerized Adaptive Test (CAT) has raised a number of concerns about its practical applications. The concerns are principally involved with the concept of having the computer construct the test from a precalibrated item pool, and substituting statistical characteristics for the test developer's skills. Problems with CAT that can arise in practice are related to context effects (item location, cross-information, unbalanced context), lack of robustness, item difficutly ordering, and high technology factors. A variety of traditional options to circumvent these problems are explored, and a new methodology for solving them is offered. In order to use the testing efficiencies of CAT and the experience of test developers, this study proposes the bundling of items into "testlets." Items in testlets can be arranged hierarchically or linearly to yield increased efficiency, but the paths an examinee can follow are strictly determined. One potential drawback of the testlet approach is that the aggregate of items is treated as a unit, and it may be difficult to replace pieces of it independently of the others. (JAZ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED279708

CATs, Testlets, and Test Construction: A Rationale for Putting Test Developers Back Into CAT

Howard Wainer and Gerard L. Kiely
ETS - T

PROGRAM STATISTICS RESEARCH

TECHNICAL REPORT NO. 86-71

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H.C. Weidenmiller
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08541

TM 870 120

BEST COPY AVAILABLE

CATs, TESTLETS, AND TEST CONSTRUCTION: A RATIONALE FOR PUTTING TEST DEVELOPERS BACK INTO CAT

Howard Wainer
Educational Testing Service

and

Gerard L. Kiely*
University of Minnesota

Program Statistics Research
Technical Report No. 86-71

Research Report No. 86-33

September 1986

* This paper is a result of work done while a summer predoctoral fellow at the Educational Testing Service, Princeton, NJ

This work was supported by the Educational Testing Service. Our thanks for helpful comments to: Robert Mislevy, Malcolm Ree, Fumiko Samejima, David Thissen, and the test development staff at ETS.

Copyright © 1986 by Educational Testing Service. All Rights Reserved.

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

ABSTRACT

It is observed that many sorts of difficulties preclude the uneventful construction of tests by a computerized algorithm, such as those currently in favor in Computerized Adaptive Testing (CAT). In this essay we discuss a number of these problems as well as some possible avenues of solution. We conclude with the development of the 'testlet', a bundle of items that can be arranged either hierarchically or linearly, which maintains the efficiency of an adaptive test while simultaneously keeping the quality control of test construction that is only possible currently with careful expert scrutiny.

TABLE OF CONTENTS

		Page
1.	Introduction	1
2.	Problems	4
	2.1 Context Effects	4
	2.2 Lack of Robustness	8
	2.3 Item Difficulty Ordering	9
	2.4 Influence of High Technology	10
3.	Traditional Solutions	11
	3.1 Context Effects	12
	3.2 Lack of Robustness	14
	3.3 Item Difficulty Ordering	15
	3.4 Influence of High Technology	17
4.	A New Model	17
	4.1 Introduction	17
	4.2 Advantages of the Testlet Model	20
	4.3 Psychometric Modeling	22
	4.3.1 Full-size Hierarchical Testlet	23
	4.3.2 Linear Testlet	24
5.	Discussion	25
	References	28
	Tables	32
	Figures	34

1. INTRODUCTION

In the modern world, computers are used for many purposes. In the field of mental testing they have long been used for scoring tests. More recently they are being used to aid in the construction of tests and even in their interpretation. Others (e.g., Green, 1986) have already commented on the problems associated with the computerized interpretation of test scores and we will not repeat those concerns here [except that for some purposes Meehl's (1954) 30 year-old arguments still sound pretty good -- especially as augmented by more recent work (Dawes and Corrigan, 1974; Wainer, 1976, 1978)]. In this paper we are concerned with the problems involved in the computerized construction of tests, especially within the context of adaptive testing.

In the old days of testing a wise examiner confronted a nervous examinee and asked questions. After getting some initial idea of the examinee's ability or knowledge level the examiner would not waste time asking questions that were to difficult nor too easy. Instead he would focus his questions on the area around the examinee's proficiency level. This made for a challenging and efficient test. It was also subjective and expensive.

This state of affairs changed with the increasing sophistication of processing recruits into the armed services where it was demonstrated that tests could make the training and accessioning process more efficient. Substituting for the wise examiner was a broad range test made up of many multiple choice questions. This had several advantages, among them was its vastly reduced cost per unit, and its objectivity -- everyone took the same test and it was scored in the same way. It also had some disadvantages. Due to its objective nature, a test form could be stolen, and because of its broad audience it had to contain items that would tax the least able of the prospective recruits as well as the

most able. Fortunately, Lincoln's observation about the abundance of the common man held and the inclusion of many items middling difficulty yielded a test that was not too inefficient for most of the examinee population.

Despite these minor drawbacks (which were dealt with in a variety of ways) the mass administered multiple choice test has enjoyed great success since the early days of the Army alpha and now tens of millions of such tests are given every year within the U.S. They enjoy similar success abroad.

The inefficiencies of mass administration become more than a minor expense if one considers the opportunity costs involved over the entire examinee population. Furthermore, a fixed-form test is not equally valid nor equally reliable over the entire ability range owing to the overabundance of items that are commonly found stacked up at the middle of the range. This was tolerated when there was no alternative, but with the widespread availability of cheap computing there arose another way; a test that would 'adapt itself' to the ability of the person taking it. Hints of this were seen earlier in Lord's (1971a,b) "Flexilevel test" which maneuvered the examinee through the test depending on his or her responses. Lord's later work (1974, 1976, 1977a, 1977b, 1980) and that of Weiss (1976, 1978, 1980) laid the groundwork for an operational computerized adaptive test (CAT).

CAT was supposed to represent a return to old fashioned testing where each person would have a test that was specifically designed for him or her and testing would continue until the person's ability was measured to the required degree of accuracy -- thus allowing all examinees to have a test of equal reliability and validity. The theoretical glue that holds the CAT together is item response theory (IRT) which was also pioneered by Lord (1952, 1980).

The basic strategy of CAT is that a person receives an item of moderate difficulty from a large pool of items and the next item presented depends upon whether or not the previous one was successfully answered. At each stage the item selection algorithm chooses the next item on the basis of which item in the pool provides the most potential information about the examinee's ability. The validity of such a scheme leans crucially on the assumption of unidimensionality of the trait being tested.

Recent experience with fledgling CATs has raised a number of serious concerns about the practical implementation of a method that theoretically at least seems eminently useable. The point of this paper is to describe the areas of our concerns and to explore some possible avenues of practice that may prove fruitful.

The validity of any test rests on the quality of its items. Items are the building blocks from which the test is made. If these are poorly made up or poorly combined, the entire enterprise is fruitless. Because of this, items are made up by subject-matter experts and are carefully edited to be as clear and as unambiguous as possible. Traditional tests are then constructed from these items so that they span a particular range of difficulty that is chosen to reflect the ability level of the examinee population. There are often many other characteristics that the items and the test must have when they are completed. The sum of these characteristics is usually referred to as the "test specifications."

In a computerized adaptive test, skilled test developers make up an item pool that satisfies the usual requirements of matching content specifications, spanning the required difficulty range, and passing a sensitivity review. These items are then calibrated (some item response model is fit to pre-test data and

each item's estimated parameters are recorded). It is at this point that the test developer's skill is no longer necessary and is replaced by the hubris of a psychometrician (commonly known as "item information").

We believe that substituting statistical characteristics for the test developer's skill in actually making up the test is, at best, premature. We feel that there are a number of other characteristics that must be considered along with statistical information before CAT, in practice, lives up to its theoretical counterpart. In section 2 of this paper we describe several problems that arise in practice that are not solved with existing methods. In section 3 we explore a variety of traditional options to circumvent these problems, and in section 4 we describe a new methodology that we believe can be employed to solve many of the problems in section 2. Section 5 summarizes our investigations.

2. PROBLEMS

2.1 Context Effects

The notion that the role of a skilled test developer can be effectively duplicated (or even improved upon!) through the use of an algorithm utilizing only a function of the item parameters, is unverified. Relying solely on item parameters does not prevent context effects.

Context effects refer to any influence or interpretation that an item may acquire purely as a result of its relationship to the other items making up a specific test. In a traditional testing situation where everyone takes the same test, context effects are the same for all examinees. And so, while they may still introduce errors, there is less concern about their differential impact.

In the case of CAT testing however, the much greater problem of differential context effects arises, since every examinee potentially takes a different

test and the context is therefore different for each examinee. Furthermore, in the CAT real-time approach to test construction, the phase of test development where context effects might ordinarily be identified and eliminated from paper-and-pencil tests, has no counterpart.

Historically, IRT has been considered immune from many of the concerns of test developers, such as context effects, through a reliance on the assumption of unidimensionality, which holds that the only factor determining responses to items of known characteristics is the common ability shared by the examinees. This assumption of unidimensionality clearly excludes the possibility of such extraneous factors as context, which is therefore presumed to be inconsequential, and argues instead in favor of the fungibility of individual test items in test construction. Fungibility refers to the characteristic of items that are equivalent in their parameters, rendering them mutually substitutable or interchangeable (i.e., their parameters are invariant and as a result, the performance of items is not altered nor distorted by their context in the test).

In spite of the importance placed upon unidimensionality in current item response theory, it has long been recognized that unidimensionality does not strictly hold in most cases. Once the absence of unidimensionality is acknowledged and the potential presence of other systematic influences on test-taking behavior is accepted, then the problems faced by test developers in the development of traditional tests, e.g., context effects, resurface.

Item Location. One undesirable item context effect is the influence that item location within a test, or in different tests, has on the item parameter estimates. Several studies (Whitely and Dawes, 1976; Yen, 1980; Eignor and Cook, 1983; Kingston and Dorans, 1984) have found differences in difficulty parameter estimates for items as a function of their location. Whitely and

Dawes (1976) found that when fifteen verbal analogy items were embedded in seven different unspeeded 60-item tests, and parameters were then obtained separately for each test, the estimates differed significantly for nine of the fifteen items. Yen (1980) found difficulty differences due to context effects for reading comprehension items administered in a nonspeeded test, with the items becoming more difficult the later in the test they appeared. A preequating study conducted by Eignor and Cook (1983) replicated Yen's findings.

In a study on the implications of item location effects for both equating and adaptive testing, Kingston and Dorans (1984) examined whether the difficulty of certain types of items is affected by exposure to the same item types earlier in the test. They found that, for two item types in particular (analysis of explanations and logical diagrams), difficulty decreased substantially as familiarity with the item format increased. These effects, they concluded, were at least partially due to the complexity and novelty of the item types. For another type of item (reading comprehension), difficulty increased moderately as the item appeared later in the test. This outcome they attributed to fatigue.

Cross-Information. Another undesirable effect of context that may appear on a test is cross-information. Cross-information refers to information that one item may inadvertently contribute about the answer to another. For example:

1. Carbon dioxide (CO_2) is added to water to make:

    a) Carbolic acid
    b) Seltzer
    c) Sloe gin fizz
    d) Hydrogen carbonate

2. CO_2 is known as:

    a) Carbon dioxide
    b) Carbon monoxide
    c) Carbonated water
    d) Ozone.

In this example, the item selection algorithm employed by CAT could select the first item shown above as the one that will provide the most information about an examinee's ability level at some point in a hypothetical test. It might then follow up with the second question shown above at some later point in the test. It can be seen that the answer to the second item is clearly contained in the first and, as a result, the item parameter values estimated for item number two in the absence of this additional information are invalid. However, the scoring algorithm, not knowing this, proceeds as if the examinee had in fact responded correctly to a more difficult item than was actually the case.

This problem of one item providing cross-information about another is an especially difficult one for CAT, because the very efficiency and score validity of CAT are reliant upon accurate estimates of the parameters which characterize each item's performance. But, if information pertaining to an item's solution is made available through another item, then all of the parameter estimates for that item become invalid: in addition to becoming easier, the probability of success through guessing may increase, and it may become less discriminating on the dimension of interest, as well.

Unbalanced Context. A third type of context effect is the repeated emphasis on a particular subject or theme throughout a test. Suppose, the selection algorithm happened to choose several items on golf. Aside from the obvious social class effect, it also increases the potential for contributory information across items.

Another type of lack of balance is in reference to gender, race, or ethnic groups across items. Such an imbalance violates current test development stan-

dards and may serve to distract examinees and create feelings of hostility or resentment that hinder their test-taking performance. An obvious example of this would be the selection of a succession of items that are limited to male gender references in professional roles.

In the paper-and-pencil test situation, a test developer can read and study each test directly during its development phase to identify and eliminate context effects. Unfortunately, CAT testing makes such a strategy difficult or impossible. At present, the only way to insure that cross-information is never inadvertently provided on a CAT test is to compare all conceivable subsets of items derivable from the item pool. This is at best an unreasonable undertaking, especially since a viable item pool must contain hundreds, if not thousands, of items and is subject to continuous updating and development. The identification and elimination of the other context effects discussed above is even more difficult under the present circumstances. It would require an analysis of the content of each and every test produced by CAT, certainly an impossibility.

2.2 Lack of Robustness

The efficiency of the CAT approach to testing allows the maintenance of a high degree of accuracy in score estimation at a significantly reduced test length. This too is not without its pitfalls. The shorter test lacks the redundancy inherent in a conventional test containing many more items, and so becomes much more vulnerable to idiosyncrasies in item performance. As a result, when an item is failing to perform in accordance with its parameter estimates on a CAT test, the detrimental impact on validity is about twice what would result from a traditional paper-and-pencil test of about the same theoretical accuracy, (but of greater length). Thus, a single instance of a flawed

item can seriously degrade the representativeness of individual response vectors. Furthermore, if the flawed item is one of moderate difficulty and relatively high discrimination, its detrimental influence can be felt many times due to CAT's tendency to favor a select subset of "choice" items.

2.3 Item Difficulty Ordering

A different kind of context effect, one that is not undesirable but rather is often planned or built into paper-and-pencil tests, is the sequential ordering of items by increasing difficulty. In the conventional testing situation, a test or section begins with items whose difficulties are less than the abilities of all but the least able of examinees, and the difficulties increase as one progresses through the test, concluding with items whose difficulty exceeds the ability of all but the most able.

Beginning a test with the easiest items and ordering them by increasing difficulty allows examinees, especially those least able, to develop confidence at the outset of the test with the easiest items. They can then work through the test to their capacity in the time allowed, without wasting time or becoming unnecessarily discouraged by items that are far too difficult for them. This approach also tends to minimize guessing since the items that those least able, (and therefore most prone to guessing) spend the greatest amount of time on, are those to which they have the greatest likelihood of knowing the answer.

In the CAT designed test however, the initial item administered is usually one of moderate difficulty for the entire group being tested. Thus, for the able examinee, the difficulty of successive items increases from this rather easy start, just as with the conventional test. For those at the low end of the ability continuum however, the test begins with items that are excessively difficult. These gradually become less difficult as incorrect responses are

recorded until, finally, the items approach a range of difficulty suitable for this group.

This approach to item difficulty ordering raises concerns about possible negative effects for low ability examinees, such as frustration and anxiety, which can hamper their performance. A number of studies that support these concerns have shown that when items are administered from most to least difficult rather than in the reverse order, the difficulty of the test increases (Mollenkopf, 1950; MacNicol, 1956; Sax and Carr, 1962; Hambleton, 1968; Monk and Stallings, 1970; Towle and Merrill, 1975). Furthermore, such an approach prolongs testing time and uses more items than necessary for both high and low ability examinees, while contributing little or no additional information about their ability levels.

2.4 Influence of High Technology

Among the problems associate with applications of sophisticated computer technology is that the technology tends to assume a life of its own, often changing or redefining the situation in the process. This tendency for computers to determine the direction of efforts can perhaps best be explained by two of the most distinctive characteristics of high technology: its cost, and the too common belief that it can do everything better in every way.

Given the tremendous financial investment involved in the application of computers to the administration of group tests, the tendency to maximize their usefulness would seem quite natural. At the same time however, the magnitude of the investment may make it difficult to recognize and adapt to any significant shortcomings arising from the implementation of this new technology. An excellent example of this can be found in the evolution of the paragraph comprehension test.

The paragraph comprehension-type of test is a highly developed format that has been in use for many years. However, despite its demonstrated usefulness, it may not be appropriate for computerized administration due to the physical requirements of the item layout. There are very few paragraph comprehension items small enough to fit entirely on a single computer screen and any manipulation of the computer required by the examinee to view the entire item may introduce a computer familiarity factor into the test. Furthermore, the conditional independence assumption of some IRT models argues against the inclusion of more than a single question accompanying each paragraph.

Rather than generating investigations into the appropriateness of computerized paragraph comprehension tests, these limitations have tended to encourage the development of shorter paragraphs, with perhaps only a single question pertaining to each. This effort to facilitate computer presentation raises concerns about whether these modified and abbreviated paragraph comprehension items are in fact measuring the same construct as originally intended. Green's (1986) factor analytic study showed that some paragraph comprehension items modified in this way look more like word knowledge items. It is, thus, a natural question to ask whether it is desirable in this instance, for the presentation medium to dictate the design or modification of item formats. How does such a change affect validity?

3. TRADITIONAL SOLUTIONS

Unfortunately, no simple solution suggests itself to the variety of problems related to CAT implementation that have been discussed. However, certain avenues of investigation invite exploration in the search for solutions.

3.1 Context Effects

One possible solution to the problems arising from context effects would be to provide additional information about individual item characteristics beyond the currently accepted item parameters. This approach is justifiable since, in the presence of context effects, factors other than the ability assumed by the model are influencing item and person performance and should therefore be accounted for.

Examples of additional item characterizations would be identifiers or codes that label each item in terms of its reference to gender, race, or national origin, that identify keywords that may suggest the presence of cross-information, or that code each item to indicate the topic or theme it represents. The CAT algorithm could then be modified to make use of this information in item selection to avoid cross-information, to balance gender, race, or ethnic references and to prevent redundancy of item topics. While this approach may improve item selection compatibility and avoid certain types of context problems, it does nothing to enhance item fungibility. Items still remain context dependent. In any case, further research is required to clarify the need for more extensive identification of item characteristics before such a course is pursued.

A disadvantage of the item characterization approach is that much larger item pools would be necessary if context effects are to be accounted for while maintaining the accuracy of measurement throughout the ability range of interest. This may lead to item pools that are too large and too expensive for practical applications. Another disadvantage is that a great deal of additional item pool development would be required. Assigning a gender or ethnic group identifier to each item is a simple matter, but identifying keywords or coding each item by topic or theme is a far more time consuming effort. Before items

can be characterized adequately, exhaustive taxonomies of principal words and topics would have to be developed. Each item would then require evaluation and classification according to these taxonomies before item pool development can be considered complete.

An alternative to the full item characterization approach would be to balance item pools in terms of the specified characteristics and then to rely on the random selection of items within difficulty and information levels to avoid or minimize any negative effects. The primary advantage of this approach is that the development of taxonomies for item characterization is not required. However, items would still need individual evaluation to identify the characteristics of interest, and item pools would require careful development to establish the desired balance of characteristics within difficulty and information levels. Furthermore, constant monitoring of item pools would be necessary to ensure that the proper balance is maintained as the item pool is updated.

Unfortunately, this type of approach does nothing to insure that problems related to item context effects will not occur, it only attempts to minimize their frequency through randomization. But in the case of a large scale testing program where millions of forms would be generated each year, every possible combination of items would likely occur. This approach also assumes that enough highly discriminating items will be available at each difficulty level that test developers will be able to select only those items with the desirable characteristics without sacrificing the measurement quality of the item pool.

A third approach to the problem of context effects is to seek refuge in the assurance that their influence is insignificant in comparison to the substantial effects of ability differences. The viability of this position has yet to be

convincingly demonstrated.

3.2 Lack of Robustness

One of the major features of CAT is that it is designed to maximize the efficiency of the testing procedure for each individual. In other words, it seeks to minimize the number of item responses required to obtain an ability estimate with some predetermined degree of accuracy. But, as test efficiency increases, robustness invariably declines. And so, the detrimental impact of a flawed item or other error on test validity increases with the brevity of the test.

The most obvious way to guard against the lack of robustness is to increase redundancy by inserting extra items into each examinee's test. But such an approach is self-defeating since it sacrifices efficiency for all examinees before testing even commences. A more rational approach would be to eliminate or contain any factors influencing examinee responses other than the ability the test is designed to measure. The development of item pools designed to protect against context effects would be a substantial contribution to this effort.

Another approach to overcoming the problem of lack of robustness is the use of a stopping rule based upon the accuracy of individual assessments. One example is a rule that continues administering items until a predetermined model-based standard error of the ability estimate is attained, at which time testing is terminated. In this way, CAT can individually correct for lack of robustness by incrementally relinquishing efficiency item-by-item until the desired accuracy of estimate is realized.

As long as the assumptions implied by the model are not violated, a model-based stopping rule will maximize measurement efficiency for each examinee. However, the kinds of problems that give rise to concerns over the lack of

robustness imply violations of the model, as well. In this instance, if an examinees' responses fail to fit the model, a model-based stopping rule may be misleading in terms of the measurement accuracy it purportedly attains. In such a case, a stopping rule that evaluates a standard error derived directly from examinee response data may be preferable.

Standard errors can be derived directly from the data by employing a resampling technique such as the bootstrap or jackknife to draw repeated samples from each examinee's responses. Ability estimates can then be obtained for each sample, and the standard error calculated from the variance of the obtained distribution of ability estimates (Wainer and Wright, 1980; Wainer and Thissen, in press). In this way, a stopping rule based on the standard error of the ability estimate can be employed to terminate testing while avoiding error due to the lack of model fit.

3.3 Item Difficulty Ordering

The problems posed for the less able examinee by the order in which item difficulties are presented by CAT are a legitimate concern. Numerous studies have shown that items of median difficulty for the individual examinee provide the most valid measurement of their ability (e.g., Thurstone, 1932; Richardson, 1936; Tucker, 1946). It would seem, therefore, that some effort to select as the initial item for each individual, one that is of moderate difficulty for them, is appropriate. This can be accomplished in a number of ways, such as using information about examinees' previous test performance, examinees' estimates of their own ability levels relative to their group, or perhaps a short routing test to identify an appropriate range of difficulty within which to commence individual testing.

Another approach to the item difficulty ordering problem would be to estimate the magnitude of the order effect for each item, and then utilize this information in adaptive test construction in the form of an additional item parameter. The magnitude of the order effects can be estimated by comparing ability estimates derived from a test administration in which difficulty order effects are absent, with those estimated from an actual adaptive test where difficulty effects are present. The difference between these two estimates can then be plotted against ability level to produce a function that describes the magnitude of the effect at each level of ability.

One way to estimate ability without order effects in a CAT-like environment is to use a "live simulation". A live simulation can be carried out by administering items similar to those utilized in the adaptive test, but in conventional paper-and-pencil form, ordered by increasing difficulty. Items would then be selected from the paper-and-pencil test and rearranged to mimic what an adaptive test for each individual would have been. An ability estimate can then be derived for each examinee based on the subset of items that would have been administered adaptively, and this estimate can be compared with that obtained from an actual adaptive test. The difficulty order effects resulting from the adaptive administration can be estimated by comparing the two measures.

The development and use of an item difficulty ordering parameter may not be practical in actual application, but the consequences of ignoring difficulty ordering may prove to be of little concern. This is because examinees in relatively close proximity to each other in ability will, on average, be subjected to approximately equal testing conditions and so their ability estimates can be fairly compared. For those further apart on the ability continuum, comparisons will be less accurate on average, but such comparisons will be less crucial and

less subject to misinterpretation, as well.

3.4 Influence of High Technology

The best defense against the powerful influence of high technology is a return to first principles in the development of items and tests to insure that construct validity is maintained. Item formats not conducive to computer presentation must be identified and, if satisfactory modifications cannot be made without altering the item's relation to the construct of interest, the format should be eliminated. Following the same line of reasoning, new item formats, ones especially favorable to computerized presentation, should be researched and developed while, at the same time, guarding against the context effects cited by Kingston and Dorans (1984) due to novel or complex item types.

4. A NEW MODEL

4.1 Introduction

The problems that arise from the CAT approach to testing may also be overcome through the application of an alternative model. One model that appears promising in this regard is a multistage fixed branching CAT model that substitutes multi-item testlets for single items as the unit for test development and administration.

A testlet is a group of items related to a single content area that is developed as a unit and which contains a fixed number of predetermined paths that an examinee may follow. In this way, each item is embedded in a predeveloped testlet, in effect carrying its own context with it. The paths through a testlet may follow a hierarchical branching scheme that routes examinees to successive items of greater or lesser difficulty depending on their previous responses, and which culminates in a series of ordered score categories (Figure 1).

Figure 1 about here

Or, the testlet may contain only a single linear path of a number of items that are administered to all examinees. Which form is chosen depends critically on the application for which it is intended (Figure 2).

Figure 2 about here

Just as branching schemes may vary within testlets, the testlets themselves may also be combined to form a complete test by linking them hierarchically or in a linear fashion, or some combination of the two, again depending on the intended purpose of the test. This arrangement allows for the construction of a wide variety of tests for specific purposes by combining hierarchical and linear branching both between and within testlets in any desired combination. To better describe this we will denote a hierarchical testlet with the symbol [<] and a linear testlet with [—] .

One application of an adaptive testlet model is the measurement of knowledge in several different content areas, such as multiplying fractions and solving word problems on a mathematics test. This type of test would consist of a separate cluster of testlets for each content area, with hierarchical branching both within and between the testlets in each cluster. The clusters of testlets would then be linearly linked so that all examinees begin at the same starting point with each new content area (Figure 3).

Figure 3 about here

With this model, testing within each content area would begin with an initial testlet designed to be of moderate difficulty for the group. This testlet would

route examinees to one of two or more follow-on testlets of greater or lesser difficulty, depending on their individual performance. In this way, examinees would be routed adaptively through and between the testlets in the cluster, at the end of which a score would be obtained for that content area before moving on to the next.

An example similar to the one above is the case where examinees are required to demonstrate minimum competency in a content area, beyond which individual performance levels are to be assessed. The test model for this situation would begin with a testlet containing a single linear path, thus administering all items in the testlet to all examinees. This testlet would be developed specifically to establish some minimum competency, and following which, examinees would be linearly routed to one or more hierarchically branched testlets designed to measure individual performance levels (Figure 4).

Figure 4 about here

Another important application of the linear testlet would be for gathering data on experimental items, perhaps for norming or equating purposes, from all examinees tested, regardless of their level of ability. It seems clear, for example, that in CAT calibration it is important to include such sections if one needs to get good estimates of item slopes. Since testlets whose items are linearly linked will administer all of the items to all examinees, response data can be gathered from all examinees, not just those for whom the items are of moderate difficulty. The testlet could be excluded from scoring, and examinees will be unaware that the algorithm is not branching them through the testlet based on their responses (Figure 5).

Figure 5 about here

4.2 Advantages of the Testlet Model

The fixed branching testlet model offers a number of significant advantages over the variable branching adaptive models currently under development. First, the testlet approach itself is adaptive, and so it enjoys most of the advantages associated with adaptive tests. However, the fixed branching feature provides only a limited number of paths for examinees to follow. Therefore, the test developer can construct each of these paths or individual tests with careful scrutiny, incorporating the desired features or characteristics into each testlet, while avoiding or neutralizing the negative influences previously identified.

When errors or blunders do occur with the testlet model, they can be much more easily overcome. Since each examinee is not subjected to a unique test that is constructed on the spot, but rather follows one of a limited number of predetermined paths, those errors that do occur will appear repeatedly. Under these circumstances it is possible to identify those examinees affected by the error, and to implement an appropriate correction for everyone in this subgroup.

Under the testlet model, item order effects are localized, and their impact is greatly diminished as a result. This is because examinees close to one another on the ability continuum will be administered nearly identical tests, and so the effects will be relatively constant within this group. For those further apart in ability, such extraneous influences are less likely to confuse their relative ability estimates and so are of much less concern.

Perhaps the greatest advantage of the testlet approach to adaptive testing lies in the implications it holds for conditional independence. The use of separate clusters of hierarchically organized testlets to assess each content area individually permits the measurement of more homogeneous abilities than

might otherwise be practical. Such homogeneity provides greater assurance that unidimensionality is maintained and conditional independence remains unviolated.

Another implication of the testlet approach for local independence follows from an important distinction made by Rosenbaum (1986). Rosenbaum distinguished between violations of latent conditional independence due solely to idiosyncratic features of test format, and actual departures from the unidimensionality of test content. Assume:

1. A test containing J dichotomously scored (1=correct, 0=incorrect) items, $X = (X_1, X_2, \ldots, X_J)$, has been administered to a random sample of examinees from a large population.
2. A unidimensional latent variable model for the population distribution of X represents pr(X=x) in terms of an unobserved scalar random variable U of the form

$$pr(X=x) = \int pr(X=x|U=u)\, dF(u)$$

where $F(\cdot)$ is the population distribution of U.
3. $I \leq J$ item bundles/testlets $(V_1, V_2, \ldots, V_I)$ are identified, and each X_j is contained in exactly one of the V_i's.
4. There is conditional independence between but not necessarily within testlets, i.e., $V_1 \perp\!\!\!\perp V_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp V_I | U$.
5. Higher scores on each testlet are expected from examinees with higher values of U, i.e., $E\{g(V_i)|U=U\}$ is nondecreasing in u for all decreasing $g(\cdot)$ and for i=1,...,I.
6. (Y_1, Y_2, Z) is any partition of X into three nonoverlapping groups of items such that each V_i is entirely contained in one of the three groups.

Then Rosenbaum's Theorem 1 states that $cov\{g_1(Y_1), g_2(Y_2)|h(Z)\} \geq 0$ for all

nondecreasing functions $g_1(\cdot)$ and $g_2(\cdot)$, and for any function $h(\cdot)$. This shows that, given the loss of conditional independence within testlets, unidimensionality can still prevail between testlets and hence, a unidimensional item response model may still be appropriate.

In the context of the testlet model, Rosenbaum's theorem permits the use of items with desirable characteristics with less concern for violation of the assumption of conditional independence. Consider a traditional paragraph comprehension item consisting of a passage of text accompanied by a series of questions about it. While it has long been acknowledged that this item type violates the conditional independence assumption, when embedded within a testlet the consequences of this violation need not overly concern the test developer.

One of the strengths of Rosenbaum's result is that, if a traditional IRT model shows a lack of fit when applied to testlet data, a statistical test using contingency table methods can be applied to determine whether the lack of fit is due to a loss of conditional independence within testlets. If this is found to be the case, then examinee performance can still be summarized with a single parameter, since the between-testlet scoring may still satisfy the requirements of a unidimensional model.

4.3 Psychometric Modeling

So far we have outlined, in general, the way that a testlet approach can aid in the solution of the variety of CAT problems elaborated upon in section 2. In this section we will be more specific, describing statistical test models and procedures that can be used to score tests developed within a testlet framework. Because the framework is general and meant to be adapted to best serve in various circumstances, so too must the analysis model. Therefore we will describe a set of models rather than a particular one.

4.3.1 Full-size hierarchical testlet

Let us begin by considering a test made up of full-size hierarchical testlets. One such testlet, depicted in Figure 6, contains three levels and seven items. The numbering of the items in the testlet reflects the ordering of the difficulty of the items (1 is the easiest item, and 7 is the most difficult). Such a testlet would yield eight groups of examinees associated

---

Figure 6 and Table 1 about here

---

with the eight possible response vectors. Note that each outcome unambiguously reflects a particular response vector and that these vectors can be strictly ordered. This is shown in Table 1. To fit such a test we could:

1. assume that conditional independence among items holds and fit the items ignoring any structure using our favorite IRT model (3-PL with priors, whose parameters are estimated via marginal maximum likelihood is our current favorite) next
2. test for goodness-of-fit (noting that -2loglikelihood is distributed as χ^2 is one way to start on this, there are other kinds of tests as well). If fit is satisfactory stop. We are done.
3. If not, check to see if lack of fit was due to grouping. If this is the case Rosenbaum's theorem holds and we can still have unidimensionality among testlets. At this point we change models and use an ordered response model (Samejima, 1969). Now each testlet yields a score from (say) 0 to 7, but the estimated θ is on the same scale and we can proceed.

4. If lack of fit was not due to grouping you do not have a unidimensional test. Consider restructuring it (perhaps using full information factor analytic approaches [Bock et al, 1985]) and trying again.

4.3.2 Linear Testlet

Suppose we have a test made up of a set of linear testlets (whether linearly or adaptively configured). Now it is possible to have within testlet response patterns that do not form a strictly ordered set. Once again we can follow the same pattern as in 4.3.1. This time if we get to step 3 we can use Andrich's DLM model (1985) which does not require local independence within testlet, and press on.

4.3.3 Mixtures

When a test is made up of testlets of various types, follow the strategies suggested above and then concatenate them (see Roche, Wainer and Thissen, 1976 for details on how to do this). The key to this concatenation is the conditional independence of testlets that can exist even in the absence of such independence within testlets (Theorem 1). Moreover, context effects should be drastically reduced since each testlet carries its own context around with it.

4.3.4 Partial Hierarchical Testlet

Suppose we have constructed a hierarchical testlet that is not full (see Figure 7). Outcomes are labeled a-f. Note that outcomes c and d are not

---

Figure 7 and Table 2 about here

---

uniquely determined; they each represent two different response vectors. These vectors are explicitly shown in Table 2; one would be hard pressed to rank order outcomes c_2 and d_1. Thus an ordered category model does not seem suitable.

Also, since the testlet is hierarchical, it is not clear that Andrich's model would work. In this instance it may be necessary to use a nominal category model (Bock, 1972; and Samejima, 1983), although its use within this context is yet to be explored.

5. DISCUSSION

In this essay we have concerned ourselves with some problems that recent attempts to develop an operational CAT have experienced. These problems are principally involved with having the computer construct the test out of a precalibrated item pool. Our conclusion was that it is impractical, given current technology, to characterize items sufficiently fully to trust that computer constructed tests will conform to the currently accepted canons of good practice. Thus, we feel it is premature to take test construction out of the hands of test developers.

To maintain the gains of testing efficiency that CAT offers while at the same time placing test developers and their accumulated wisdom back into the test construction loop, we proposed the bundling of items into 'testlets'. These groups of items can be hierarchically designed and so yield increased efficiency, but the paths an examinee can follow are strictly determined. Thus a test developer can take pains to avoid the context effects elaborated upon previously.

We view the testlet as a rather flexible unit, but one useful size testlet would be one of four levels (containing 15 items, partitioning examinees into 16 classes). A full hierarchical testlet of more than four levels yields too many different paths for easy scrutiny; less than four levels may be suitable for some applications, but could not be used where subscore estimates are required. If current research on the increased efficiency of CAT holds, such a testlet

would have roughly the same amount of information about an examinee as six or seven items administered in a linear way. This matches rather well with most modern reading comprehension tests (e.g., the College Board's Degrees of Reading Power Test) which contain a passage and then about seven items. We feel this would be an admirable way to employ testlets.

One potential drawback of the testlet approach is that the aggregate of items is treated as a unit, and so it may be difficult to replace pieces of it independently of the others. This is particularly true if an ordered category model is used in scoring. We suspect that sometimes this will be possible, sometimes not. One might view this as an inefficient use of items, since the first item in a hierarchical testlet will be used very often (and so need replacement relatively quickly) whereas the last items are used much less frequently (and last much longer). Sometimes this will cause us anguish (anyone who has ever had to junk a car whose body was rotted out while its engine was still good understands this perfectly). Yet in many cases this will not be a problem. For example, consider paragraph comprehension -- it would be very difficult to make up more than 15 items on a single moderate length passage, but even if it could be done, surely the ones made up later (items 16+) would not be as good as the earlier ones. The best questions come first; later item writers have to stretch hard. Thus replacing passage and items when usage requires it, is not as profligate as might first appear.

Our position of abandoning computerized construction of tests within a CAT framework may seem radical on the surface, but it is not. We believe in adaptive testing as a methodology, but we feel there is ample evidence that the Fisherian information derived from the parameters of currently used test models is simply insufficient to effectively construct tests. We are advocating the

inclusion of more information into the item choice algorithm. Since at the moment we know of no way to concatenate all of the item's characteristics into such an information measure we propose instead placing human test developers back into the process. This assumption, that test developers do and know something important, seems like a rather conservative approach to building tests. Thus, our advocacy of not allowing an algorithm to build the test unaided until the algorithm contains as much wisdom as the expert test developer is not radical — it is merely sensible. Hierarchical testlets are a scheme which can maintain the CAT advantages while still using the wisdom of experts.

REFERENCES

Andrich, D. (1985). A latent trait model for items with response dependencies: Implications for test construction and analysis. In S. E. Embretson (Ed.), Test Design. New York: Academic Press.

Bock, R. D. (1972) Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Bock, R. D., Gibbons, R. D., and Muraki, E. (1985). Full-information item factor analysis. (MRC Report No. 85-1). Chicago, IL: National Opinion Research Center.

Dawes, R. M. and Corrigan, B. (1974). Linear models in decision making. Psychological Bulletin, 81, 95-106.

Eignor, D. R. and Cook, L. L. (1983). An investigation of the feasibility of using item response theory in the preequating of aptitude tests. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Green, B. F. (1987). Construct validity of computer-based tests. In H. Wainer and H. Braun (Eds.), Test Validity. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K. (1968). Effects of item order and anxiety on test performance and stress. Paper presented to Division D, American Educational Research Association annual meeting, Chicago, February.

Kingston, N. M. and Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. Applied Psychological Measurement, 8, 146-154.

Lord, F. M. (1952). A theory of test scores. Psychometric Monograph, No. 7.

Lord, F. M. (1971a). Robbins-Monro procedures for tailored testing. Educational and Psychological Measurement, 31, 3-31.

Lord, F. M. (1971b). The self-scoring flexilevel test. Journal of Educational Measurement, 8, 147-151.

Lord, F. M. (1974). Individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, and P. Suppes (Eds.), Contemporary Developments in Mathematical Psychology, Vol. II. San Francisco, Freeman.

Lord, F. M. (1976). Some likelihood functions found in tailored testing. In C. L. Clark (Ed.), Proceedings of the First Conference on Computerized Adaptive Testing. Washington, D.C.: United States Civil Service Commission.

Lord, F. M. (1977a). A broad-range tailored test of verbal ability. Applied Psychological Measurement, 1, 95-100.

Lord, F. M. (1977b). Practical applications of item characteristic curve theory. Journal of Educational Measurement, 14, 117-138.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

MacNicol, K. (1956). Effects of varying order of item difficulty in an unspeeded verbal test. Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press.

Mollenkopf, W. G. (1950). An experimental study of the effects on item analysis data of changing item placement and test-time limit. Psychometrika, 15, 291-315.

Monk, J. J. and Stallings, W. M. (1970). Effect of item order on test scores. Journal of Educational Research, 63, 463-465.

Richardson, M. W. (1936). The relation between the difficulty and the differential validity of a test. Psychometrika, 1, 33-49.

Roche, A., Wainer, H., and Thissen, D. (1976). Skeletal Maturity. New York: Plenum Publishing.

Rosenbaum, P. R. (1986). A note on item bundles. Princeton, NJ: Educational Testing Service. (Unpublished manuscript -- submitted to Psychometrika).

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph, No. 17.

Samejima, F. (1983). A latent trait model for differential strategies in cognitive performance. Office of Naval Research Report, 83-1.

Sax, G., and Carr, A. (1962). An investigation of response sets on altered parallel forms. Educational and Psychological Measurement, 22, 371-376.

Thurstone, T. G. (1932). The difficulty of a test and its diagnostic value. Journal of Educational Psuchology, 23, 335-343.

Towle, N. J. and Merrill, P. F. (1975). Effects of anxiety type and item difficulty sequencing on mathematics test performance. Journal of Educational Measurement, 12, 241-249.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. Psychometrika, 11, 1-13.

Wainer, H. (1976). Estimating coefficients in linear models: it don't make no nevermind. Psychological Bulletin, 83, 213-217.

Wainer, H. (1978). On the sensitivity of regression and regressors. Psychological Bulletin, 85, 267-273.

Wainer, H. and Thissen, D. M. (198). Estimating ability with the wrong model. Journal of Educational Statistics, (in press).

Wainer, H. and Wright, B. D. (1980). Robust estimation of ability in the Rasch model. Psychometrika, 45, 373-391.

Weiss, D. J. (1976). Adaptive testing research at Minnesota: Overview, recent results, and future directions. In C. L. Clark (Ed.), Proceedings of the First Conference on Computerized Adaptive Testing. Washington, D.C.: United States Civil Service Commission.

Weiss, D. J., (Ed.) (1978). Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota.

Whitely, S. E. and Dawis, R. V. (1976). The influence of test context on item difficulty. Educational and Psychological Measurement, 36, 329-337.

Yen, W. M. (1980). The extent, causes, and importance of context effects on item parameters for two latent trait models. Journal of Educational Measurement, 17, 297-311.

TABLE 1

Pattern	Item 1	2	3	4	5	6	7	Score
a	0	0		0				0
b	1	0		0				1
c		1	0	0				2
d		1	1	0				3
e				1	0	0		4
f				1	1	0		5
g				1		1	0	6
h				1		1	1	7

TABLE 2

	Item						
Pattern	1	2	3	4	5	6	Score
a	0	0		0			0
b	1	0		0			1
c_1		1	0	0			2
c_2			0	1	0		2 or 3 or 4?
d_1		1	1	0			3 or 4?
d_2			1	1	0		5
e				1	1	0	6
f				1	1	1	7

Figure 1

A 3 Level, 7 Item, 8 Outcome, Fully Hierarchical Testlet

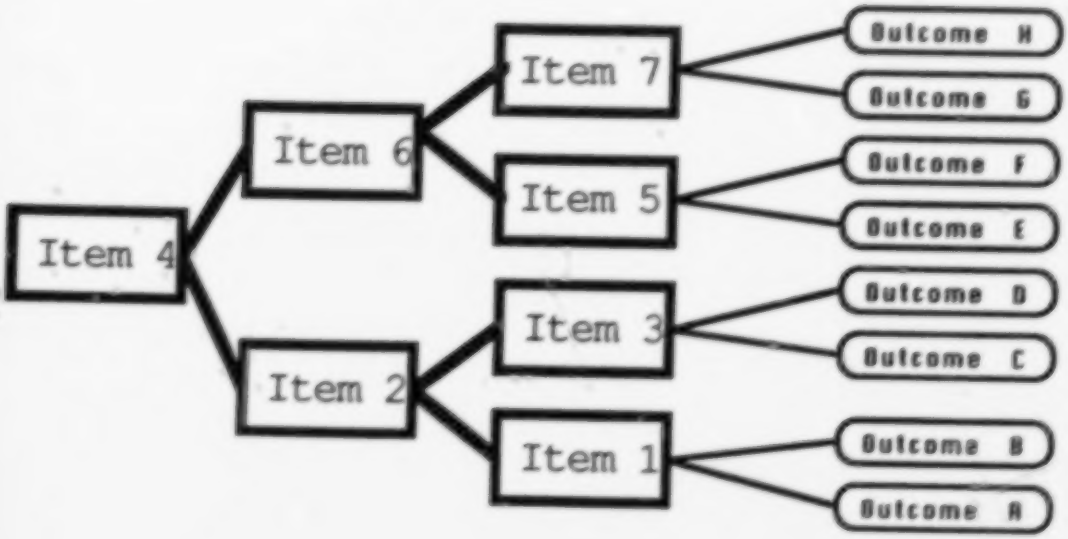


Figure 2

A 3 Level, 3 Item Linear Testlet

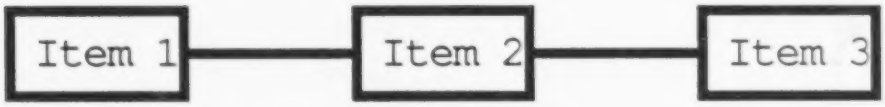


Figure 3

A Test Design Showing Two Test Areas Linearly Joined Where Each Area is Measured with Hierarchically Connected Hierarchical Testlets

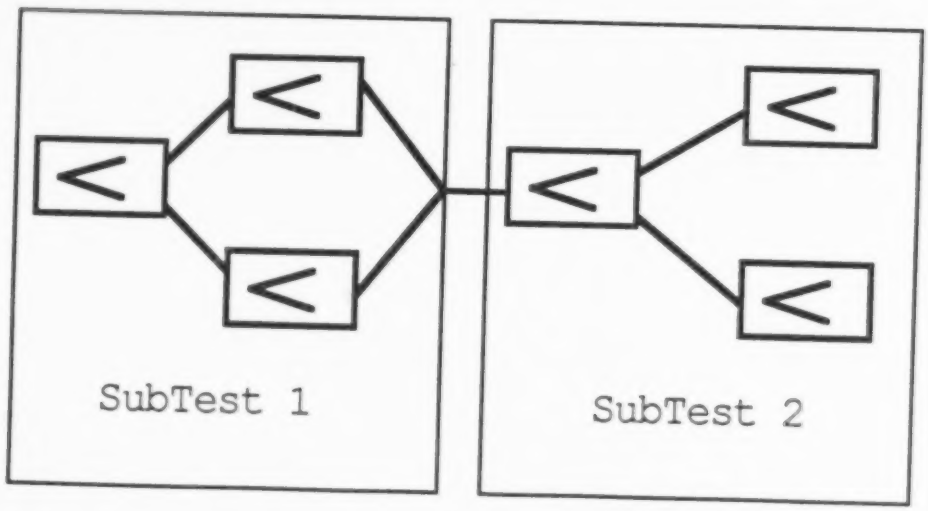


Figure 4

A Test Design Showing Two Test Areas Joined Where An Identical First Test is Given To Everyone And Branching Test Is Given Thereafter

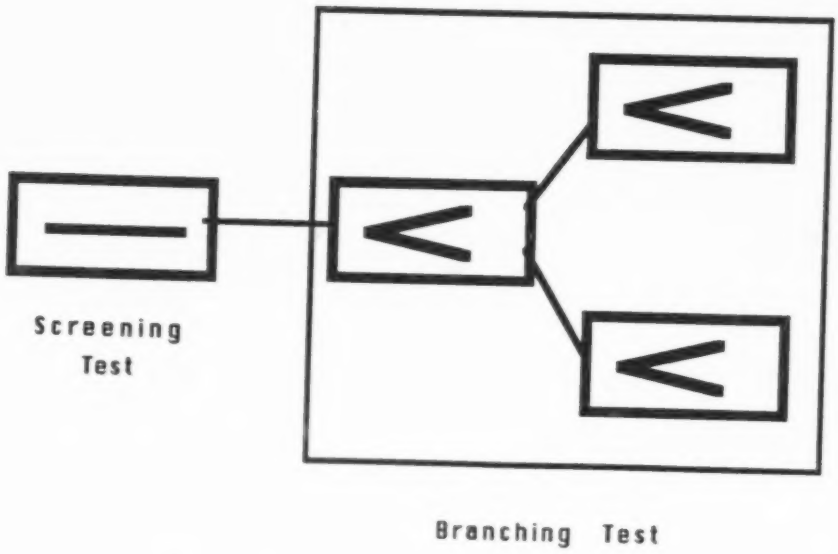


Figure 5

A Hierarchical Test Design
With An Experimental Testlet

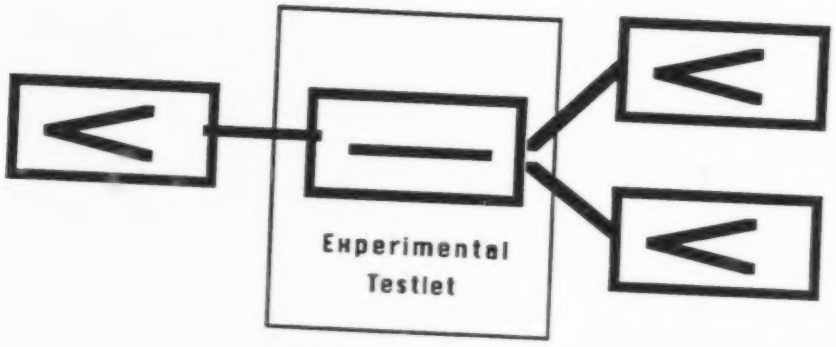


Figure 6

A 3 Level, 7 Item, 8 Outcome, Fully Hierarchical Testlet

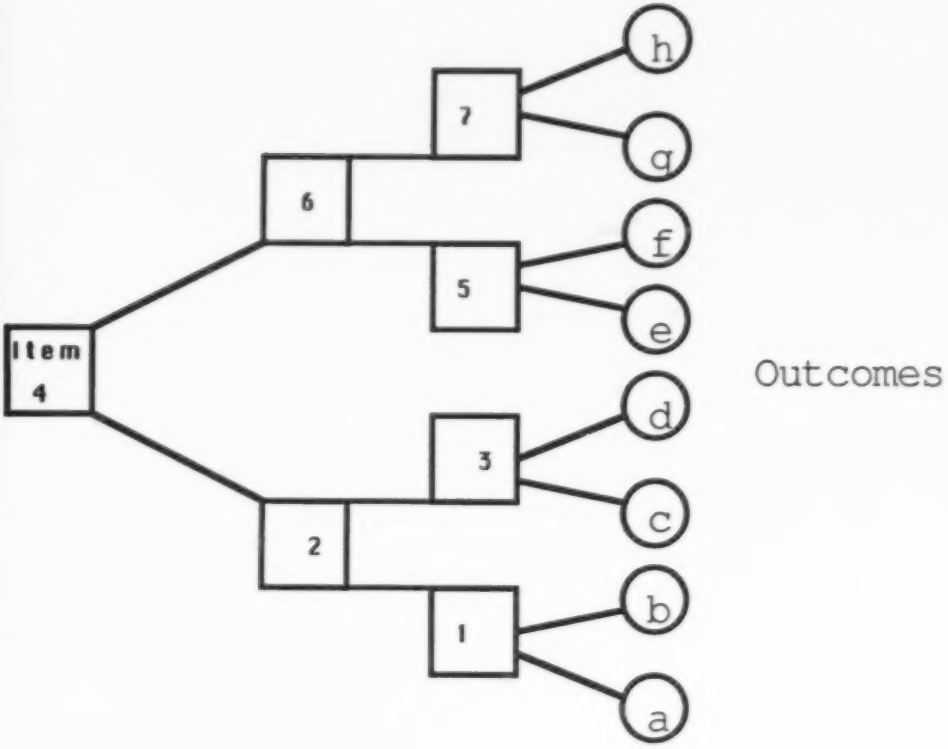
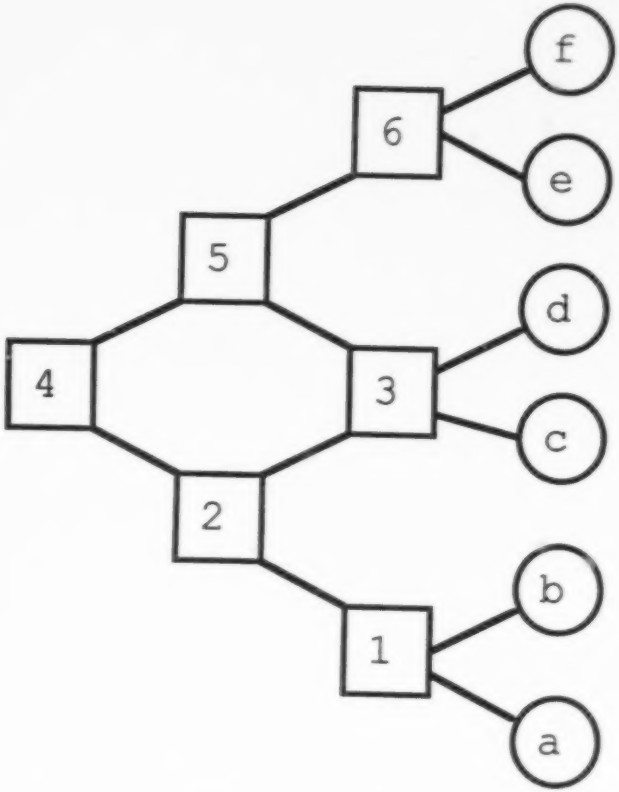


Figure 7

A 3 Level, 6 Item, 6 Outcome Partially Hierarchical Testlet



Outcomes